Jonathan Skaggs

CS 478 Nearest Neighbor

Magic Telescope (without distance weighting)

without normalization (k=3):

Training set accuracy: 0.879715072041, Test set accuracy: 0.808280828083

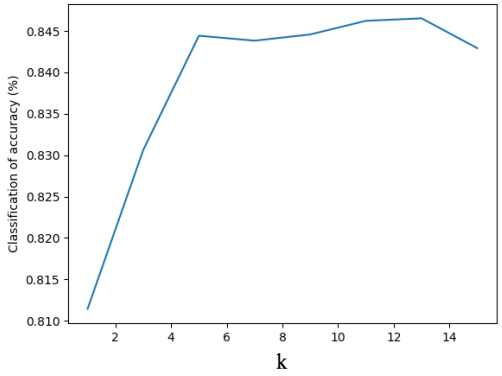
with normalization (k=3):

Training set accuracy: 0.89630888781, Test set accuracy: 0.830633063306

We can see that this improvement was because we are looking at all features rather than being bias toward large value features in the dataset.

Test Accuracy

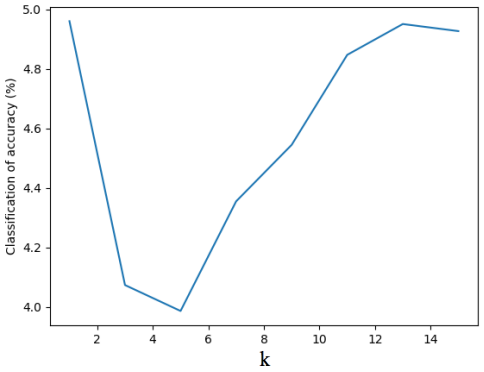
1: 0.811431143114
3: 0.830633063306
5: 0.844434443444
7: 0.843834383438
9: 0.844584458446
11: 0.846234623462
13: 0.846534653465
15: 0.842934293429



Housing Price Prediction (without distance weighting)

Mean Sq. Error

1: 4.96068859863
3: 4.07414933581
5: 3.98710175342
7: 4.35538698146
9: 4.54520264701
11: 4.84757088088
13: 4.95108618768
15: 4.92729174309



Magic Telescope (with distance weighting)

without normalization (k=3):

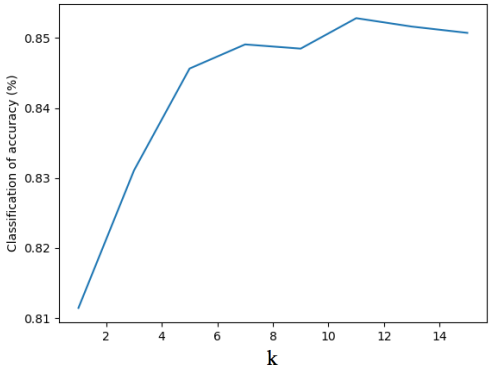
Training set accuracy: 1.0, Test set accuracy: 0.808580858086

with normalization (k=3):

Training set accuracy: 1.0, Test set accuracy: 0.831083108311

Test Accuracy

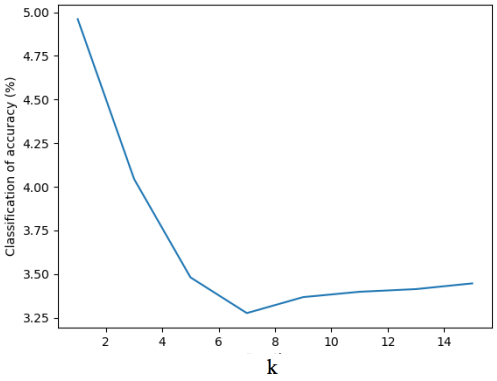
1: 0.811431143114
3: 0.831083108311
5: 0.845634563456
7: 0.849084908491
9: 0.848484848485
11: 0.852835283528
13: 0.851635163516
15: 0.850735073507



Housing Price Prediction (with distance weighting)

Mean Sq. Error

1: 4.96068859863
3: 4.04489851615
5: 3.48132555098
7: 3.27662439109
9: 3.36824181432
11: 3.39875161823
13: 3.41406529665
15: 3.4468418368



Analysis

We can see that, using distance weighting helps improve accuracy for both nominal and continuous data. In some cases, the improvement was minimal but overall there was significant improvement. There was improvement because, closer data tends to be more correct. It is especially important as k increases.

Credit Approval

Distance Metric

For continuous data, I used regression just as in the last two experiments. For nominal data, anything that was not 0 was 1.

Unknown Data

To handle unknown data, I decided that I would rather trust in data that I had. If the data was unknown I assumed that feature was far away. I manually changed the specific feature to be 1 unit away (after normalization).

Results (k=3)

Training set accuracy: 0.842857142857, Test set accuracy: 0.785

More Experiments

I decided to try to weight each feature with by the amount of information gained, similar to what we did in decision tree. My hypothesis is that on the credit score data set, using information gain to weight features, will increase the accuracy. Because calculating information gain does not work well with continuous data I don't think that the magic and housing dataset will perform any differently. I tested on the housing and magic dataset and I was correct. The accuracies were essentially stayed the same with only a .0001 difference. I was a little surprised, because on critic dataset when weighting by information gain, there was a slightly worse accuracy. I think if I was going to continue to run experiments I would try to classify continuous data into sets of nominal data. I think it might work better.

Critic dataset

Training set accuracy: 0.822448979592, Test set accuracy: 0.78